Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Accelerated oblique random survival forests

Byron C. Jaeger BJAEGER@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Sawyer Welden SWELDEN@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Kristin Lenoir KLENOIR@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Jaime L Speiser JSPEISER@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Matthew Segar MATTHEW.SEGAR@UTSOUTHWESTERN.EDU
Division of Cardiology, Department of Internal Medicine,
University of Texas Southwestern Medical Center, Dallas

Nicholas M. Pajewski NPAJEWSK@WAKEHEALTH.EDU
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA

Editor: TBD

Abstract

The oblique random survival forest (ORSF) is an ensemble method for supervised learning that extends the random survival forest (RSF). Trees in the ORSF are grown using linear combinations of variables to create branches in the tree, whereas in the RSF a single variable is used. ORSF ensembles often have higher prediction accuracy than RSF ensembles, but the additional computational overhead of fitting ORSF ensembles limits their scope of application. In addition, few methods have been developed for interpretation of ORSF ensembles. In this article, we introduce and evaluate methods to accelerate the ORSF (that is, reduce computational overhead) and compute the importance of individual variables in the ORSF We show that our strategy to accelerate the ORSF is up to 500 times faster than existing software for ORSFs (the `obliqueRSF` R package), and that prediction accuracy of the accelerated ORSF is equivalent or superior to that of existing ORSF methods. We estimate importance of variables for the ORSF by negating each coefficient used for the given variable in linear combinations, and then computing the reduction in out-of-bag accuracy. We show with simulation that 'negation importance' can discriminate between signal and noise variables, and it outperforms several state-of-the-art variable importance techniques in this task when there is correlation among predictors.

Keywords: Random Forests, Survival, Efficient, Variable Importance

1. Introduction

Risk prediction can reduce the burden of disease by educating patients and providers and guiding strategies to prevent and treat disease in a wide range of medical domains (Moons et al., 2012a,b). The random survival forest (RSF), a supervised learning algorithm that can engage with censored outcomes, is frequently used for risk prediction. Notable characteristics of the RSF include uniform convergence of its ensemble survival function to the true population survival function when the predictor space is discrete (Ishwaran and Kogalur, 2010). In addition, software implementing the RSF is freely available, extremely efficient, and full of tools to interpret and explain the RSF (Ishwaran and Kogalur, 2019; Wright and Ziegler, 2017; Hothorn et al., 2010). However, there remains considerable potential to improve the RSF in risk prediction tasks where training samples are not large enough to guarantee asymptotic properties or predictor spaces are non-discrete (that is, predictors are continuous).

RSFs may be axis based or oblique. The axis based RSF uses a single predictor whereas the oblique RSF uses a linear combination of predictors to create branches in trees. While axis based decision boundaries are always perpendicular to the axis of the relevant predictor, linear combinations of predictors create oblique decision boundaries that are neither parallel nor perpendicular to axes of their contributing predictors. Prior work has found the oblique RSF has higher prediction accuracy than the axis based RSF in general benchmarks (Jaeger et al., 2019) and that oblique splitting is particularly effective when predictors are continuous (Menze et al., 2011). However, existing methods to implement oblique splitting typically use fully trained models in each non-leaf node to identify linear combinations

of predictors, exponentially increasing the number of operations required for the oblique RSF versus its axis based counterpart. In addition, standard methods to estimate variable importance (VI) in the RSF are less effective in the oblique RSF, and few methods have been introduced to estimate VI specifically for the oblique RSF.

The aim of this article is to improve the computational efficiency and interpretability of the oblique RSF. In a general benchmark experiment including 31 risk prediction tasks, we show that oblique RSFs with partially trained models have equivalent or superior prediction accuracy and are orders of magnitude more efficient than oblique RSFs with fully trained models in non-leaf nodes. We introduce a method to estimate VI for oblique RSFs and compare its ability to discriminate between signal and noise variables versus standard and state-of-the-art methods. All methods proposed in this article are available in the `aorsf` R Package.

2. Related work

2.1 Axis-based and oblique random forests

After Breiman (2001) introduced the axis-based and oblique random forest (RF), numerous methods were developed to grow oblique RFs for classification or regression tasks (Menze et al., 2011; Zhang and Suganthan, 2014; Rainforth and Wood, 2015; Zhu et al., 2015; Poona et al., 2016; Qiu et al., 2017; Tomita et al., 2020; Katuwal et al., 2020). However, oblique splitting approaches for classification or regression may not generalize to survival tasks (for example, see Zhu, 2013, Section 4.5.1), and most research involving the RSF has focused on forests with axis-based trees (Wang and Li, 2017).

Building on prior research for bagging survival trees (Hothorn et al., 2004), Hothorn et al. (2006) developed an axis-based RSF in their framework for unbiased recursive partitioning, more commonly referred to as the conditional inference forest (CIF). Zhou et al. (2016) developed a rotation forest based on the CIF and Wang and Zhou (2017) developed a method for extending the predictor space of the CIF. Ishwaran et al. (2008) developed an axis-based RSF with strict adherence to the rules for growing trees proposed in Breiman (2001). Jaeger et al. (2019) developed the oblique RSF following the bootstrapping approach described in Breiman's original RF and incorporating early stopping rules from the CIF.

2.2 Variable importance

Breiman (2001) introduced permutation VI, defined for each predictor as the difference in a RF's estimated generalization error before versus after the predictor's values are randomly permuted. Strobl et al. (2007) identified bias in permutation VI driven by variable selection bias and effects induced by bootstrap sampling, and proposed an unbiased permutation VI based on unbiased recursive partitioning (see Hothorn et al. (2006)). Menze et al. (2011) introduced an approach to estimate VI for oblique RFs that computes an analysis of

variance (ANOVA) table in non-leaf nodes to obtain p-values for each predictor contributing to the node. The ANOVA VI[1] is then defined for each predictor as the number of times a p-value associated with the predictor is ≤ 0.01 while growing a forest. Lundberg and Lee (2017) introduced a method to estimate VI using SHapley Additive exPlanation (SHAP) values, which estimate the contribution of a predictor to a model's prediction for a given observation. SHAP VI is computed for each predictor by taking the mean absolute value of SHAP values for that predictor across all observations in a given set.

3. The acceleracted oblique random survival forest

Consider the usual framework for survival analysis with training data

$$\mathcal{D}_{\text{train}} = \{(T_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^{N_{\text{train}}}.$$

Here, T_i is the event time if $\delta_i = 1$ and last point of contact if $\delta_i = 0$, and $\boldsymbol{x}_i$ is a vector of predictors values. Assuming there are no ties, let $t_1 < \ldots < t_m$ denote the m unique event times in $\mathcal{D}_{\text{train}}$.

To accelerate the oblique RSF, we propose to identify linear combinations of predictor variables in non-leaf nodes by applying Newton Raphson scoring to the partial likelihood function of the Cox regression model:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{\boldsymbol{x}_{j(i)}^T \boldsymbol{\beta}}}{\sum_{j \in R_i} e^{\boldsymbol{x}_j^T \boldsymbol{\beta}}}, \tag{1}$$

where R_i is the set of indices, j, with $T_j \geq t_i$ (i.e., those still at risk at time t_i), and $j(i)$ is the index of the observation for which an event occurred at time t_i. Newton Raphson scoring is an extremely fast estimation procedure, and the `survival` package includes documentation that outlines how to efficiently program it (Therneau, 2022). Briefly, a vector of estimated regression coefficients, $\hat{\beta}$, is updated in each step of the procedure based on its first derivative, $U(\hat{\beta})$, and second derivative, $H(\hat{\beta})$:

$$\hat{\beta}^{k+1} = \hat{\beta}^k + U(\hat{\beta} = \hat{\beta}^k)\, H^{-1}(\hat{\beta} = \hat{\beta}^k)$$

For statistical inference, it is recommended to complete iterations of Newton Raphson scoring until a convergence threshold is met. However, to identify a valid linear combination of predictors, only one iteration of Newton Raphson scoring is needed. It was not clear to us whether completing only one iteration of Newton Raphson scoring would provide coefficients that were stable enough to partition data effectively when we developed this technique, so we analyzed three learners in numerical experiments that assessed prediction accuracy (see Section 4.1):

1. Menze et al. (2011) name their method 'oblique RF VI', but we use the name 'ANOVA VI' in this article to avoid confusing Menze's approach with other approaches to estimate VI for oblique RFs.

- `aorsf-extratrees`: Randomly select coefficients for linear combinations of predictors from a uniform distribution.

- `aorsf-fast`: Identify coefficients for linear combinations of predictor variables using one iteration of Newton Raphson scoring.

- `aorsf-cph`: Identify linear combinations of predictor variables using the coefficients derived from a Cox proportional hazards model. That is, use the coefficient estimates derived from completing iterations of Newton Raphson scoring until a convergence threshold is met or the number of iterations completed is 15, whichever occurs first.

3.1 Negation variable importance

Negation VI is similar to permutation VI in that it measures how much a model's prediction error increases when a variable's role in the model is de-stabilized. More specifically, negation VI measures the increase in an oblique RF's prediction error after flipping the sign of all coefficients linked to a variable (that is, negating them). As the magnitude of a coefficient increases, so does the probability that negating it will change the oblique RF's predictions. Since the coefficients in each non-leaf node of an oblique RFs are adjusted for the accompanying predictors, negation VI may provide better estimation of VI in the presence of correlated variables compared to standard VI techniques.

For consistency with prior VI techniques for the RSF and for its computational efficiency, we use Harrell's concordance (C)-statistic (Harrell et al., 1982) to measure change in prediction error when computing negation VI. We also note that while the current article focuses on oblique RSFs, negation VI can be applied to any oblique RF and can be applied with any applicable error function.

4. Numeric experiments

4.1 Benchmark of prediction accuracy and computational efficiency

The aim of this numeric experiment is to evaluate and compare the accelerated oblique RSF with its predecessor (the oblique RSF from the `obliqueRSF` R package) and with other machine learning algorithms for risk prediction. Inferences drawn from this experiment include equivalence and inferiority tests based on Bayesian linear mixed models.

4.1.1 Learners

We consider five classes of learners: axis based RSFs, oblique RSFs, boosting ensembles, regression models, and neural networks (Table 1). For each class, we synchronized shared tuning parameters. For example, for RSF learners, we set the minimum node size (a parameter shared by all RSF learners) as 10. Additionally, for RSF learners, the number of randomly selected predictors was the square root of the total number of predictors

rounded to the nearest integer, and the number of trees in the ensemble was 500. For learners that required tuning (that is, boosting, regression, and neural networks), nested 10-fold cross-validation was applied to tune relevant model parameters. Specifically, tuning for boosting models included identifying the number of steps to complete. For regression models, tuning was used to identify the magnitude of penalization. For neural networks, the number and density of layers was tuned.

Learner Class	Software	Learners	Description
Random Survival Forests			

Axis based	`RandomForestSRC`
	`ranger`
	`party`
	`rotsf`
	`rsfse`
		`rsf-standard`
		`rsf-extratrees`
		`cif-standard`
		`cif-rotate`
		`cif-spacextend`
			`rsf-standard` grows survival trees following Leo Breiman's original random forest algorithm with variables and cut-points selected to maximize a log-rank statistic. `rsf-extratrees` grows survival trees with randomly selected features and cut-points. `cif-standard` uses the framework of conditional inference to grow survival trees. `cif-rotate` extends `cif-standard` by applying principal component analysis to random subsets of data prior to growing each survival tree. `cif-spacextend` derives new predictors for each tree in the ensemble, separately.

Oblique	`obliqueRSF`
	`aorsf`
		`obliqueRSF-net`
		`aorsf-net`
		`aorsf-fast`
		`aorsf-cph`
		`aorsf-extratrees`
			Oblique survival trees following Leo Breiman's random forest algorithm. Linear combinations of inputs are derived using `glmnet` in `obliqueRSF-net` and `aorsf-net`, using Newton Raphson scoring for the Cox partial likelihood function in `aorsf-fast` and `aorsf-cph`, and chosen randomly from a uniform distribution in `aorsf-extratrees`. Cut-points are selected to maximize a log-rank statistic.

Boosting ensembles			
Trees	`xgboost`	`xgboost-cox` `xgboost-aft`	`xgboost-cox` maximizes the Cox partial likelihood function, whereas `xgboost-aft` maximizes the accelerated failure time likelihood function. Nested cross validation (5 folds) is applied to tune the number of trees grown, the minimum number of observations in a leaf node was 10, the maximum depth of trees was 6, and $\sqrt{p}$ variables were considered randomly for each tree split, where p is the total number of predictors.
Regression models			
Cox Net	`glmnet`	`glmnet-cox`	The Cox proportional hazards model is fit using an elastic net penalty. Nested cross validation (5 folds) is applied to tune penalty terms.
Neural networks			
Cox Time	`survivalmodels`	`nn-cox`	A neural network based on the proportional hazards model with time-varying effects. Nested cross-validation was applied to select the number of layers (from 1 to 8), the number of nodes in each layer (from $\sqrt{p}/2$ to $\sqrt{p}$), and the number of epochs to complete (up to 500). A drop-out rate of 10% was applied during training.

Table 1: Learning algorithms assessed in numeric studies

4.1.2 Evaluation of prediction accuracy

Our primary metric for evaluating the accuracy of predicted risk is the integrated and scaled Brier score (Graf et al., 1999). Consider a testing data set:

$$\mathcal{D}_{\text{test}} = \{(T_i, \delta_i, x_i)\}_{i=1}^{N_{\text{test}}}.$$

Let $\widehat{S}(t \mid x_i)$ be the predicted probability of survival up to a given prediction horizon of $t > 0$. For observation i in $\mathcal{D}_{\text{test}}$, let $\widehat{S}(t \mid \boldsymbol{x}_i)$ be the predicted probability of survival up to a given prediction horizon of $t > 0$. Define

$$\widehat{\text{BS}}(t) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \{\widehat{S}(t \mid \boldsymbol{x}_i)^2 \cdot I(T_i \leq t, \delta_i = 1) \cdot \widehat{G}(T_i)^{-1} + [1 - \widehat{S}(t \mid \boldsymbol{x}_i)]^2 \cdot I(T_i > t) \cdot \widehat{G}(t)^{-1}\}$$

where $\widehat{G}(t)$ is the Kaplan-Meier estimate of the censoring distribution. As $\widehat{\text{BS}}(t)$ is time dependent, integration over time provides a summary measure of performance over a range of plausible prediction horizons. The integrated $\widehat{\text{BS}}(t)$ is defined as

$$\widehat{\mathcal{BS}}(t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \widehat{\text{BS}}(t) dt. \tag{2}$$

In our results, t_1 and t_2 are the 25th and 75th percentile of event times, respectively. $\widehat{\mathcal{BS}}(t_1, t_2)$, a sum of squared prediction errors, can be scaled to produce a measure of explained residual variation (that is, an R^2 statistic) by computing

$$R^2 = 1 - \frac{\widehat{\mathcal{BS}}(t_1, t_2)}{\widehat{\mathcal{BS}}_0(t_1, t_2)} \tag{3}$$

where $\widehat{\mathcal{BS}}_0(t_1, t_2)$ is the integrated Brier score when a Kaplan-Meier estimate for survival based on the training data is used as the survival prediction function $\widehat{S}(t)$. We refer to this R^2 statistic as the index of prediction accuracy (IPA) (Kattan and Gerds, 2018).

Our secondary metric for evaluating predicted risk is the time-dependent concordance (C)-statistic. We compute the first time-dependent C-statistic proposed by Blanche et al. (2013, Equation 3), which is interpreted as the probability that a risk prediction model will assign higher risk to a case (that is, an observation with $T \leq t$ and $\delta = 1$) versus a non-case (that is, an observation with $T > t$). Similar to the IPA, observations with $T \leq t$ and $\delta = 0$ only contribute to inverse proportion of censoring weights for the time-dependent C-statistic.

Both the IPA and time-dependent C-statistic generally take values between 0 and 1. To avoid presenting an excessive amount of leading zeroes in our tables, figures, and text, we scale both the IPA and time-dependent C-statistic by 100. For example, we present a value of 25 if the IPA is 0.25, 87 if the time-dependent C-statistic is 0.87, and present 10.2 if the difference between two IPA values is 0.102

4.1.3 Data sets

We use a collection of 18 publicly available data sets to benchmark the prediction accuracy and computational efficiency of the accelerated ORSF and each of the other learners described in Section 4.1.1. The number of right-censored outcomes per data set ranged from one to four, and the total number of risk prediction tasks we analyzed was 31 (Table A.1). Across all prediction tasks, the number of observations ranged from 137 to 17,549 (median: 2,231), the number of predictors ranged from 7 to 1,692 (median: 41), and the percentage of censored observations ranged from 5.26 to 97.7 (median: 78.2).

4.1.4 Monte-Carlo cross validation

For each risk prediction task, we completed 10 runs of Monte-Carlo cross validation. In each run, we used a random sample containing 50% of the available data for training and the remaining 50% for testing each of the learners described in Section 4.1.1. Then, for each learner, we computed the IPA, time-dependent C-statistic, and computational time required to fit a prediction model and compute risk predictions. If any learner failed to obtain predictions on any particular split of data[2], the results for that split were omitted from downstream analyses.

4.1.5 Statistical analysis

After collecting data from 10 replications of Monte-Carlo cross validation for all 14 learners in all 31 risk prediction tasks, we analyzed the resulting 4,340 observations of IPA and, separately, time-dependent C-statistic, using a Bayesian linear mixed model. Our approach follows the ideas described by Benavoli et al. (2017) and Kuhn and Wickham (2020), who developed guidelines on making statistical comparisons between learners using resampling and Bayesian methods. Specifically, we fit two models:

$$\text{IPA} = \widehat{\gamma}_0 + \widehat{\gamma} \cdot \text{learner} + (1 \,|\, \text{data/run})$$

and

$$\text{C-stat} = \widehat{\gamma}_0 + \widehat{\gamma} \cdot \text{learner} + (1 \,|\, \text{data/run}).$$

Random intercepts for specific splits of data (that is, `run` in the model formula) were nested within datasets. The intercept, $\widehat{\gamma}_0$, was the expected value using `aorsf-fast` so that the coefficients in $\widehat{gamma}$ could be interpreted as expected differences between `aorsf-fast` and other learners. Default priors from `rstanarm` were applied for model fitting (Goodrich et al., 2022).

2. For example, when the prediction task was to predict risk of death in the ACTG 320 clinical trial (26 events total), some splits did not leave enough events in the training data to fit complex learners such as the neural network

Hypothesis testing For both the IPA and time-dependent C-statistic, we conducted equivalence and inferiority tests based on a 1 point region of practical equivalence. More specifically, we concluded that two learners had practically equivalent IPA or time-dependent C-statistic if there was a 95% or higher posterior probability that the absolute difference in the relevant metric was less than 1. We concluded that one learner was weakly superior when there was ≥ 0.95 posterior probability that the difference in the relevant metric was non-zero, and concluded superiority when when there was ≥ 0.95 posterior probability that the difference in the relevant metric was 1 or more.

4.1.6 Results

Index of prediction accuracy Compared to learners that were not oblique RSFs, `aorsf-fast` had the highest IPA in out of 31 risk prediction tasks, with an overall mean IPA of (Figure 1). Compared to the learner with the second highest mean IPA (`rsf-standard`), `aorsf-fast`'s mean was 1.21 points higher, a relative increase of 9.98%. The posterior probability of `aorsf-fast` and `aorsf-cph` having practically equivalent expected IPA was 0.97, and the posterior probability of `aorsf-fast` having a superior IPA to other learners ranged from 0.65 (versus `rsf-standard`) to >0.999 (versus several other learners; see Figure 2)

Time-dependent concordance statistic Compared to learners that were not oblique RSFs, `aorsf-fast` had the highest time-dependent C-statistic in out of 31 risk prediction tasks, with an overall mean of (Figure 3). Compared to the learner with the second highest mean C-statistic (`obliqueRSF-net`), `aorsf-fast`'s mean was 0.681 points higher, a relative increase of 0.887%. The posterior probability of `aorsf-fast` and `aorsf-cph` having practically equivalent expected time-dependent C-statistics was 0.99, and the posterior probability of `aorsf-fast` having a superior time-dependent C-statistic versus other learners ranged from 0.01 (versus `obliqueRSF-net`) to >0.999 (versus several other learners; see Figure 4)

Computational efficiency Overall, `aorsf-fast` was the second fastest learner, with an expected model development and risk prediction time about 1/2 second longer than `glmnet-cox` (Figure 5). Compared to its predecessor, `obliqueRSF-net`, `aorsf-fast` was XYZ times faster.

4.2 Benchmark of variable importance

The aim of this experiment is to evaluate negation VI and similar VI methods based on how well they can discriminate between variables that do or do not have a relationship with a simulated outcome. We consider methods that are intrinsic to the oblique RF (for exampleANOVA VI), those that are intrinsic to the RF (for examplepermutation VI), and those that are model-agnostic (for exampleSHAP VI).

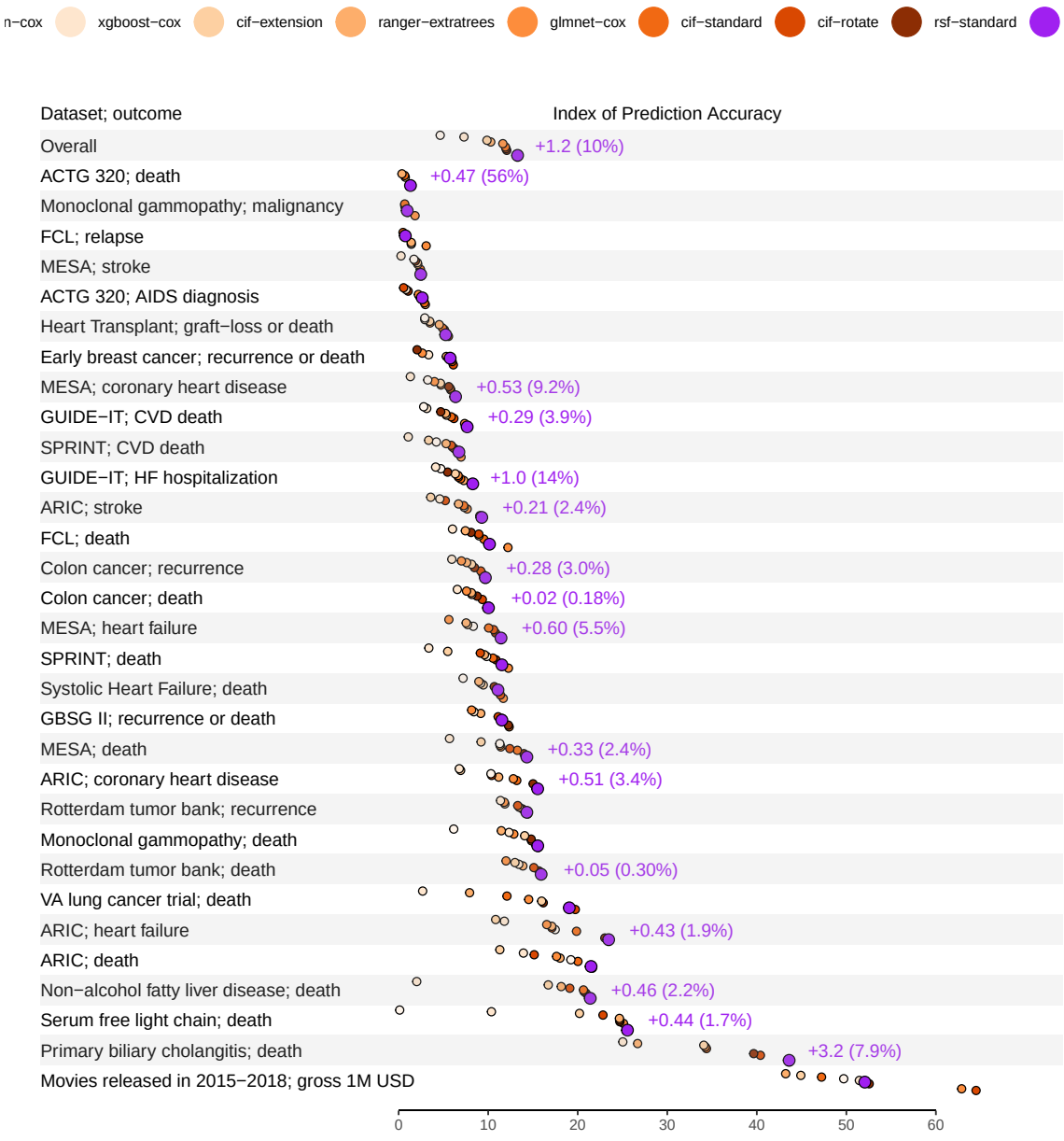

Figure 1: Index of prediction accuracy for the accelerated oblique random survival forest and other learning algorithms across multiple risk prediction tasks. Text appears in tasks where the accelerated oblique random survival forest obtained the highest index of prediction accuracy, showing the absolute and percent improvement over the second best learner.

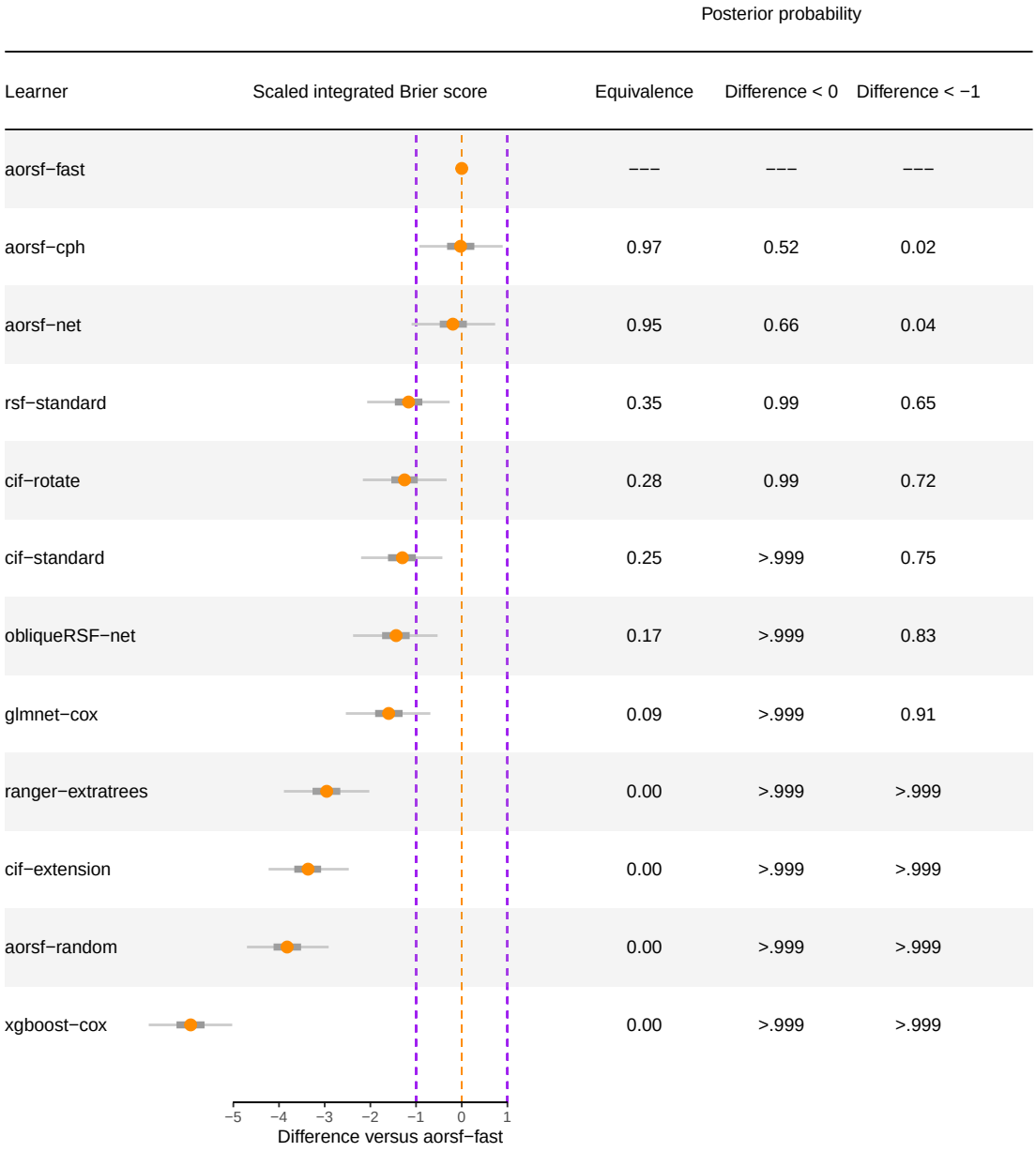

Figure 2: Expected differences in index of prediction accuracy between the accelerated oblique random survival forest and other learning algorithms. A region of practical equivalence is shown by purple dotted lines, and a boundary of non-zero difference is shown by an orange dotted line at the origin.

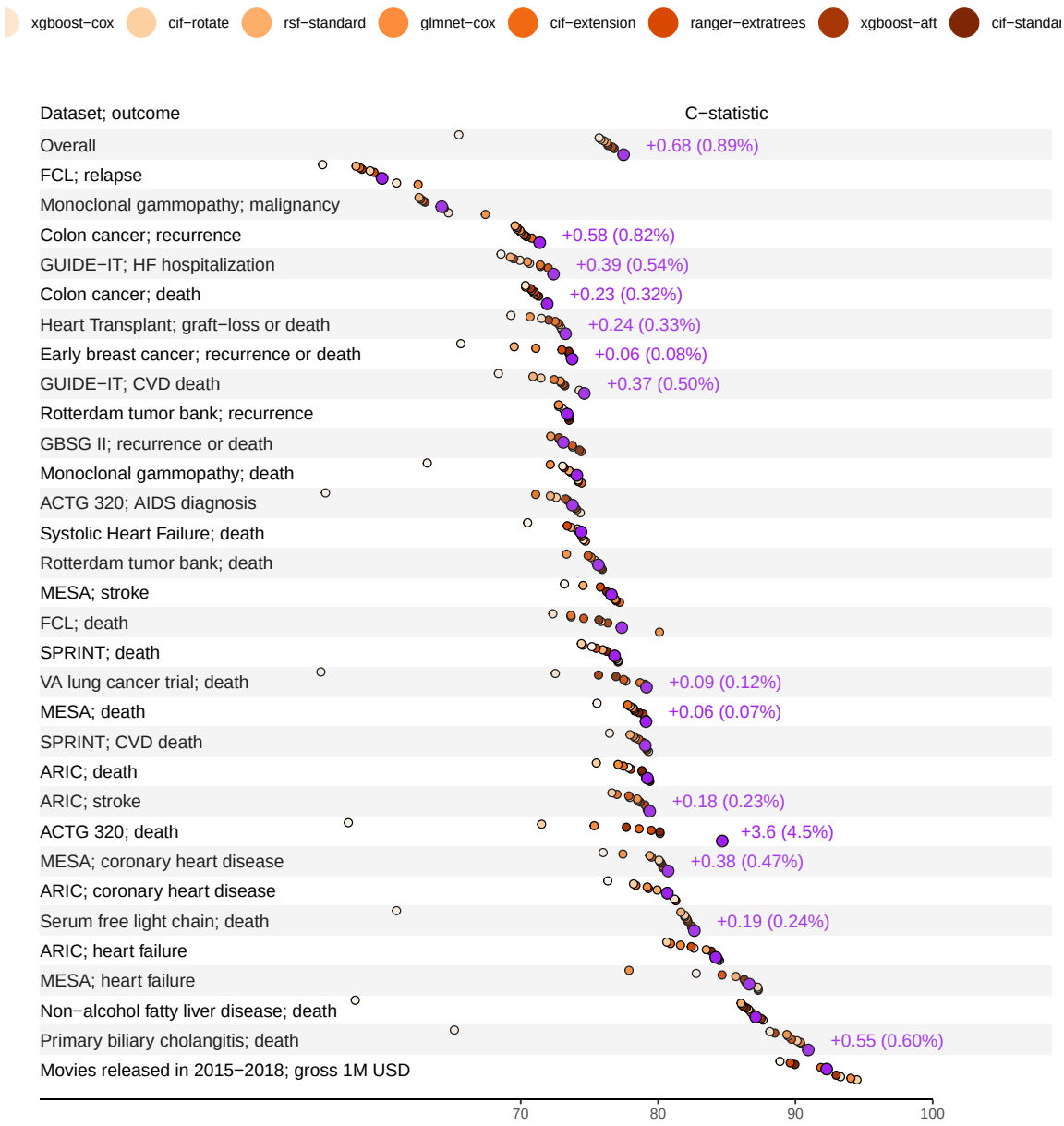

Figure 3: Time-dependent concordance statistic for the accelerated oblique random survival forest and other learning algorithms across multiple risk prediction tasks. Text appears in tasks where the accelerated oblique random survival forest obtained the highest concordance, showing the absolute and percent improvement over the second best learner.

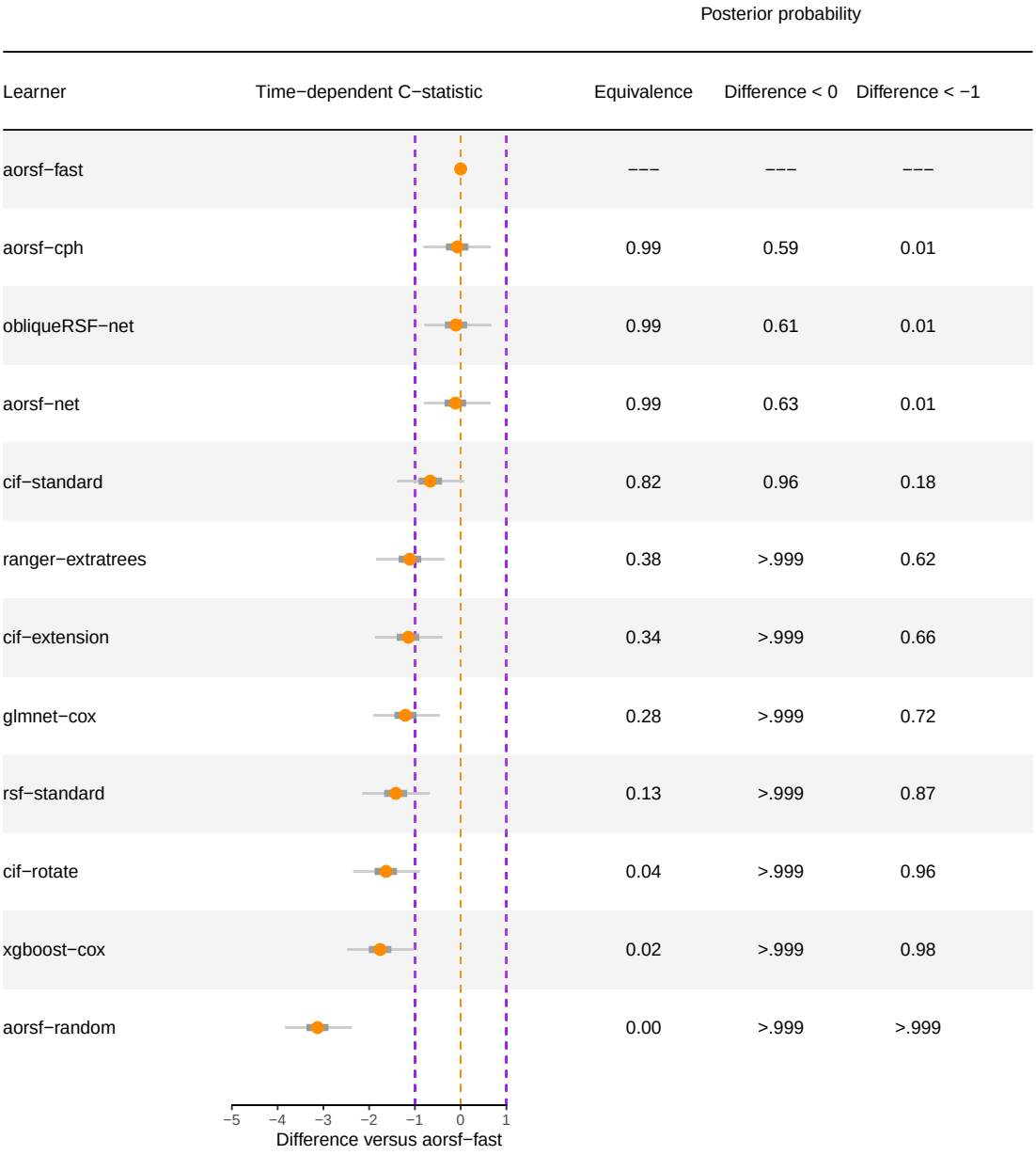

Figure 4: Expected differences in time-dependent concordance statistic between the accelerated oblique random survival forest and other learning algorithms. A region of practical equivalence is shown by purple dotted lines, and a boundary of non-zero difference is shown by an orange dotted line at the origin.

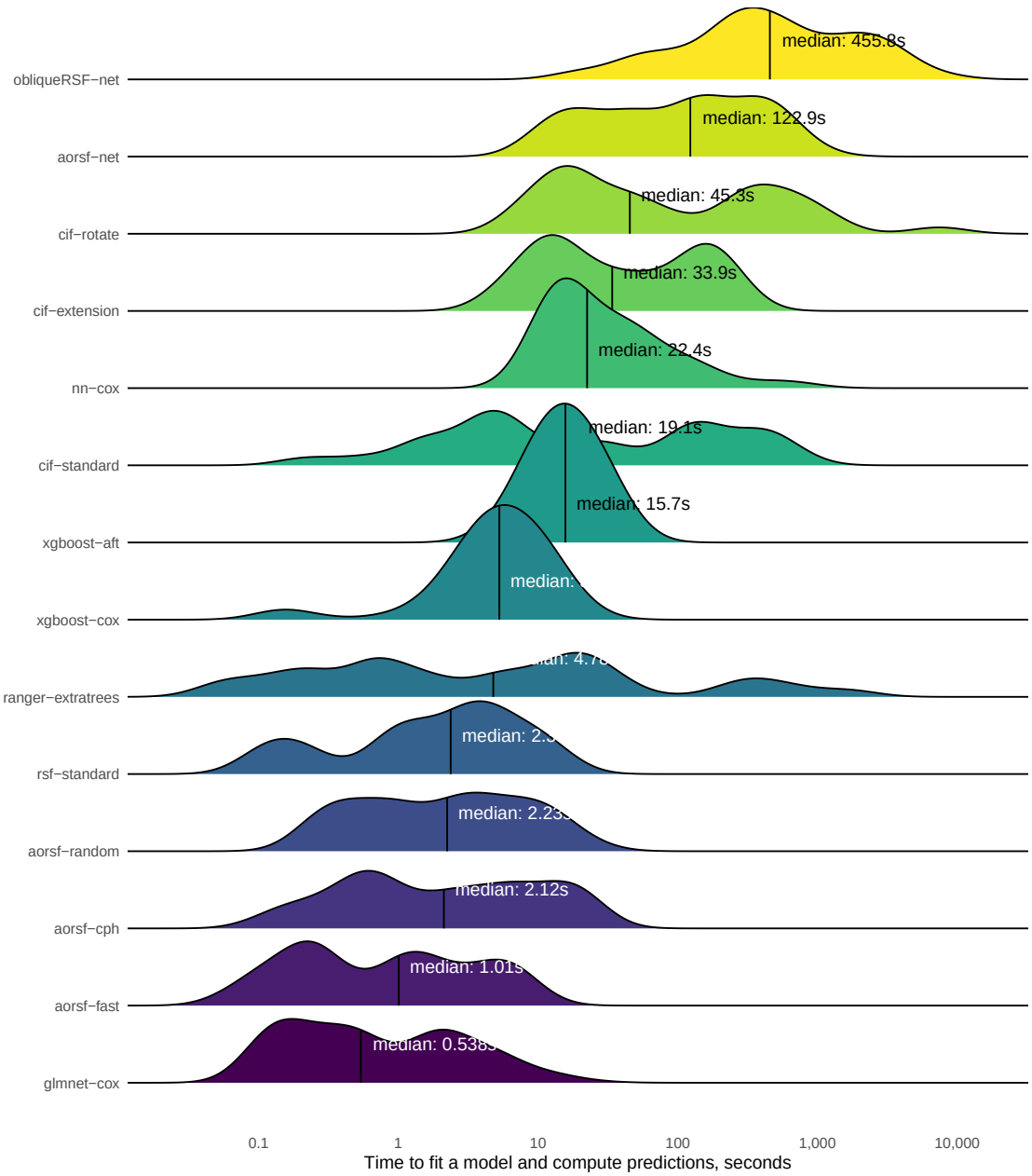

Figure 5: Distribution of time taken to fit a prediction model and compute predicted risk. The median time, in seconds, is printed and annotated for each learner by a vertical line.

4.2.1 Variable importance techniques

We compute permutation VI for axis based RSFs using the `randomForestSRC` package. Although the `party` package implements the approach to VI developed by Strobl et al. (2007), the developers of the `party` package note that the implementation of this approach for survival outcomes is "extremely slow and experimental" as of version 1.3.10. Therefore, it is not incorporated in the current simulation study. We compute ANOVA VI, negation VI, and permutation VI for oblique RSFs using the `aorsf` package. For ANOVA VI, we applied a p-value threshold of 0.01, following the threshold recommended by Menze et al. (2011). We compute SHAP VI for boosted tree models using the `xgboost` package, which incorporates the tree SHAP approach proposed by Lundberg et al. (2018). We also compute SHAP VI for accelerated oblique RSFs using the `fastshap` package.

4.2.2 Variable types

We considered five classes of predictor variables, with each class characterized by its variables' relationship to a right-censored outcome. Specifically,

- *irrelevant* variables had no relationship with the outcome.
- *main effect* variables had a linear relationship to the outcome.
- *non-linear effect* variables had a non-linear relationship to the outcome.
- *combination effect* variables were formed by linear combinations of three other variables. While their combination was linearly related to the outcome, each of the three variables contributing to the combination had no relation to the outcome.
- *interaction effect* variables were related to the outcome by multiplicative interaction with one other variable, which could have been a main effect, non-linear effect, or combination effect variable.

4.2.3 Simulated data

We initiated each set of simulated data with a random draw of size n from a p-dimensional multivariate normal distribution, yielding n observations of p predictors. Each of p predictor variables had a mean of zero, standard deviation of 1, and correlation with other predictor variables drawn at random between a lower and upper boundary. A time-to-event outcome with roughly 45% of observations censored was generated using the `simsurv` package. The full predictor matrix (that is, including interactions, non-linear mappings, and combinations) was used to generate the outcome. Interactions, non-linear mappings, and combinations were dropped from the predictor matrix after the outcome was generated so that VI techniques could be valuated based on their ability to detect these effects.

4.2.4 Parameter specifications

Parameters that varied in the current simulation study included the number of observations (1000, 3000, and 5000) and the minimum and maximum correlation between predictors (-0.1 to 0.1, -0.3 to 0.3, and -0.5 to 0.5). Parameters that remain fixed throughout the study included the number of predictors in each class (15) and the effect size of each predictor, with an increase of one standard deviation associated with a 64% increase in relative risk.

4.2.5 Evaluation of variable importance

We compared VI techniques based on their discrimination (that isC-statistic) between relevant and irrelevant variables. Specifically, we generated a binary outcome for each predictor variable based on its relevance (that is, the binary outcome is 1 if the variable is relevant, 0 otherwise). Treating VI as if it were a 'prediction' for these binary outcomes yields a C-statistic is interpreted as the probability that the VI technique will rank a relevant variable higher than an irrelevant variable (Harrell et al., 1982).

4.2.6 Results

5. Discussion

In this paper, we have developed two contributions to the oblique RSF: (1) the accelerated oblique RSF (that is, `aorsf-fast`) and (2) negation VI. Our technique to accelerate the oblique RSF reduces the number of operations required to find linear combinations of inputs using a single iteration of Newton Raphson scoring, while our VI technique directly engages with coefficients in linear combinations of inputs to measure importance of individual variables. In numeric experiments, we found that that `aorsf-fast` is orders of magnitude more efficient and just as accurate in risk prediction tasks compared to its predecessor, `obliqueRSF-net`. We also found several cases where negation VI allows oblique RSF models to discriminate between relevant and irrelevant variables more effectively than three standard methods to estimate VI: permutation, ANOVA, and SHAP VI. Our results favored negation VI in scenarios where oblique RSFs were the underlying model used to compute VI and also in scenarios where other modeling techniques (for example, boosted tree ensembles) were used.

5.1 Implications of our results

Accurate risk prediction models have the potential to improve healthcare by directing timely interventions to patients who are most likely to benefit. However, prediction models that cannot be interpreted and explained have no place in clinical practice. The current study advances the oblique RSF, an accurate risk prediction model, several steps towards being an accurate *and* interpretable risk prediction model. The improved computational efficiency of the accelerated oblique RSF increases the feasibility of applying model-agnostic

methods (for example, SHAP values) for interpretation. Faster model evaluation and refitting also improve diagnosis and resolution of model-based issues (for example, model calibration deteriorates over time). The introduction of negation VI also advances interpretability. VI is intrinsically linked to model fairness, as it can be used to identify when protected characteristics such as race, religion, and sexuality are inadvertently used (either directly or through correlates of these characteristics) by a prediction model. Since negation VI engages with the coefficients used in linear combinations of variables, a major component of oblique RSFs, it may be more capable of diagnosing unfairness in oblique RSFs compared to permutation importance and model-agnostic VI techniques.[3] Three additional consequences of engaging with coefficients in linear combinations of predictors is that negation VI

1. is non-random and hence reproducible without setting a random seed (a limitation of permutation importance)

2. does not incur bias from permutation of variables into unlikely or impossible combinations (a limitation of permutation importance).

3. can be applied to any type of oblique RF (a limitation of ANOVA importance, which can only be applied when p-values are calculable within nodes).

These characteristics suggest negation VI can be developed into a general tool for interpretation of oblique RFs.

5.2 Limitations and next steps

The current study has several limitations. The accelerated oblique RSF does not account for competing risks, and so our benchmark of prediction accuracy divided tasks where competing risks were present into event-specific tasks. Biased estimation of incidence may occur when competing risks are ignored, and allowing the oblique RSF to account for competing risks is a high priority for future studies. In addition, missing data are not addressed in the accelerated oblique RSF, and users are expected to impute missing values before passing training or testing data into exported functions from the `aorsf` R package. However, missing data are common and it is standard for ensemble tree methods to handle missing data during the tree growing procedure. Thus, a second item of high priority for future studies is to develop and evaluate strategies to handle missing data while growing an oblique RSF.

Acknowledgments

3. In our numeric experiments, negation VI's out-performance of other methods was most pronounced in smaller samples, giving some evidence supporting this hypothesis.

Research reported in this publication was supported by the Center for Biomedical Informatics, Wake Forest University School of Medicine. The project described was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR001420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Appendix

A.1: Data sets used for numeric experiments

Label	N observations	N predictors	Outcome	N Events	% Censored
VA lung cancer trial	137	8	Death	128	6.57
Colon cancer	929	12	Recurrence	468	49.6
			Death	452	51.3
Primary biliary cholangitis	276	19	Death	111	59.8
Movies released in 2015-2018	551	46	Gross 1M USD	522	5.26
GBSG II	686	10	Recurrence Or Death	299	56.4
Systolic Heart Failure	2,231	41	Death	726	67.5
Serum free light chain	7,874	10	Death	2,169	72.5
Non-alcohol fatty liver disease	17,549	24	Death	1,364	92.2
Rotterdam tumor bank	2,982	11	Recurrence	1,518	49.1
			Death	1,272	57.3
ACTG 320	1,151	12	AIDS Diagnosis	96	91.7
			Death	26	97.7
GUIDE-IT	894	59	Cardiovascular Death	110	87.7
			Hf Hospitalization	288	67.8
Early breast cancer	614	1,692	Recurrence Or Death	134	78.2
SPRINT	9,361	174	Cardiovascular Death	521	94.4
			Death	1,644	82.4
Heart Transplant	3,787	52	Graft-Loss Or Death	500	86.8
			Death	76	86.0

FCL	541	7	Relapse	272	49.7
Monoclonal gammopathy	1,384	8	Death	963	30.4
			Malignancy	115	91.7
MESA	6,783	48	Heart Failure	339	95.0
			Coronary Heart Disease	439	93.5
			Stroke	292	95.7
			Death	1,297	80.9
ARIC	13,623	41	Heart Failure	2,981	78.1
			Coronary Heart Disease	2,282	83.2
			Stroke	1,323	90.3
			Death	6,662	51.1

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks.

	Performance metric (SD)		Computation time, seconds	
	Scaled Brier	C-Statistic	Model fitting	Risk prediction
Overall				
aorsf-fast	0.133 (0.114)	0.775 (0.072)	0.801	0.170
aorsf-cph	0.132 (0.114)	0.774 (0.071)	1.965	0.173
aorsf-net	0.131 (0.116)	0.774 (0.071)	122.683	0.166
rsf-standard	0.121 (0.117)	0.761 (0.075)	2.089	0.252
cif-rotate	0.120 (0.130)	0.758 (0.081)	39.689	8.754
cif-standard	0.119 (0.100)	0.768 (0.070)	4.254	15.360
obliqueRSF-net	0.118 (0.086)	0.774 (0.071)	376.210	24.600
glmnet-cox	0.116 (0.123)	0.763 (0.074)	0.532	0.004
ranger-extratrees	0.103 (0.088)	0.764 (0.068)	2.519	2.159
cif-extension	0.099 (0.095)	0.763 (0.072)	23.746	8.318
aorsf-random	0.094 (0.084)	0.743 (0.064)	2.088	0.178
xgboost-cox	0.073 (0.102)	0.757 (0.088)	5.270	0.005
nn-cox	0.046 (0.106)	0.655 (0.132)	16.511	5.207
xgboost-aft	-4.78 (4.97)	0.767 (0.076)	15.673	0.007
ACTG 320; AIDS diagnosis, n = 1151, p = 12				
aorsf-random	0.031 (0.015)	0.739 (0.031)	0.396	0.036
ranger-extratrees	0.030 (0.014)	0.734 (0.033)	0.448	0.146
cif-standard	0.029 (0.028)	0.740 (0.043)	0.988	4.459
obliqueRSF-net	0.027 (0.022)	0.735 (0.039)	26.729	15.372
aorsf-cph	0.026 (0.024)	0.742 (0.040)	0.496	0.036
aorsf-fast	0.026 (0.022)	0.738 (0.040)	0.161	0.036
cif-extension	0.023 (0.015)	0.711 (0.041)	9.394	4.051
glmnet-cox	0.022 (0.020)	0.741 (0.024)	0.182	0.003
aorsf-net	0.021 (0.030)	0.735 (0.038)	18.601	0.037
rsf-standard	0.011 (0.030)	0.722 (0.044)	0.149	0.065
xgboost-cox	0.008 (0.033)	0.743 (0.035)	4.586	0.003
cif-rotate	0.006 (0.036)	0.726 (0.035)	15.373	3.791
nn-cox	-0.008 (0.016)	0.558 (0.129)	13.641	0.549
xgboost-aft	-7.76 (1.18)	0.733 (0.029)	13.603	0.007
ACTG 320; death, n = 1151, p = 12				
aorsf-fast	0.013 (0.022)	0.847 (0.054)	0.104	0.022
aorsf-cph	0.012 (0.021)	0.833 (0.061)	0.335	0.021

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
aorsf-net	0.010 (0.032)	0.829 (0.062)	13.672	0.024
obliqueRSF-net	0.009 (0.013)	0.839 (0.052)	8.926	10.688
aorsf-random	0.008 (0.016)	0.810 (0.069)	0.218	0.023
cif-extension	0.007 (0.014)	0.786 (0.050)	8.178	3.578
cif-standard	0.007 (0.019)	0.801 (0.057)	1.008	4.449
ranger-extratrees	0.004 (0.018)	0.795 (0.062)	0.423	0.266
xgboost-cox	-0.004 (0.003)	0.500 (0.000)	0.145	0.002
nn-cox	-0.004 (0.004)	0.574 (0.111)	14.113	0.560
rsf-standard	-0.007 (0.048)	0.801 (0.059)	0.090	0.036
cif-rotate	-0.026 (0.034)	0.715 (0.093)	13.721	3.326
glmnet-cox	-0.027 (0.072)	0.753 (0.101)	0.257	0.002
xgboost-aft	-22.8 (5.25)	0.777 (0.078)	11.919	0.006
ARIC; coronary heart disease, n = 13623, p = 41				
aorsf-fast	0.155 (0.007)	0.807 (0.005)	4.481	1.501
aorsf-net	0.152 (0.006)	0.807 (0.004)	498.111	1.449
aorsf-cph	0.151 (0.006)	0.806 (0.004)	15.754	1.558
rsf-standard	0.150 (0.007)	0.799 (0.004)	11.130	1.219
obliqueRSF-net	0.143 (0.004)	0.809 (0.004)	2875.869	334.594
cif-standard	0.132 (0.003)	0.807 (0.004)	70.329	357.765
glmnet-cox	0.128 (0.014)	0.792 (0.007)	1.606	0.019
ranger-extratrees	0.112 (0.003)	0.793 (0.005)	398.270	131.923
cif-rotate	0.104 (0.004)	0.782 (0.007)	550.670	69.312
nn-cox	0.103 (0.039)	0.763 (0.093)	40.780	85.754
aorsf-random	0.097 (0.005)	0.769 (0.006)	11.207	1.357
cif-extension	0.069 (0.002)	0.784 (0.005)	173.969	56.767
xgboost-cox	0.068 (0.017)	0.812 (0.003)	9.465	0.016
xgboost-aft	-4.83 (0.210)	0.813 (0.003)	33.629	0.014
ARIC; death, n = 13623, p = 41				
aorsf-net	0.217 (0.007)	0.792 (0.004)	940.117	2.346
rsf-standard	0.216 (0.006)	0.789 (0.004)	13.902	1.221
aorsf-fast	0.215 (0.008)	0.792 (0.004)	6.921	2.306
aorsf-cph	0.215 (0.007)	0.791 (0.004)	23.219	2.393
obliqueRSF-net	0.206 (0.005)	0.790 (0.004)	7672.274	317.021
cif-standard	0.200 (0.005)	0.788 (0.004)	69.166	373.964
nn-cox	0.192 (0.007)	0.778 (0.004)	302.175	153.845

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
ranger-extratrees	0.180 (0.006)	0.780 (0.005)	678.904	94.798
glmnet-cox	0.176 (0.019)	0.771 (0.009)	2.425	0.018
cif-rotate	0.151 (0.008)	0.755 (0.007)	613.021	69.650
xgboost-cox	0.139 (0.015)	0.794 (0.005)	17.317	0.018
aorsf-random	0.130 (0.005)	0.735 (0.005)	21.763	2.126
cif-extension	0.113 (0.002)	0.775 (0.005)	177.598	52.953
xgboost-aft	-1.56 (0.099)	0.794 (0.005)	25.317	0.013
ARIC; heart failure, n = 13623, p = 41				
aorsf-fast	0.235 (0.007)	0.842 (0.006)	5.652	1.828
rsf-standard	0.230 (0.007)	0.835 (0.005)	9.050	1.075
aorsf-cph	0.230 (0.007)	0.841 (0.006)	17.921	2.008
aorsf-net	0.229 (0.007)	0.841 (0.006)	613.168	1.685
obliqueRSF-net	0.213 (0.006)	0.841 (0.006)	3954.453	334.564
cif-standard	0.199 (0.005)	0.839 (0.006)	70.715	362.530
nn-cox	0.175 (0.025)	0.826 (0.006)	138.006	231.741
cif-rotate	0.171 (0.004)	0.806 (0.005)	562.939	67.797
ranger-extratrees	0.171 (0.004)	0.824 (0.006)	1532.153	144.888
glmnet-cox	0.165 (0.046)	0.816 (0.018)	2.192	0.012
aorsf-random	0.140 (0.005)	0.790 (0.008)	15.028	1.599
xgboost-cox	0.118 (0.016)	0.845 (0.005)	14.769	0.019
cif-extension	0.108 (0.003)	0.809 (0.007)	195.201	60.594
xgboost-aft	-3.80 (0.291)	0.844 (0.005)	38.814	0.015
ARIC; stroke, n = 13623, p = 41				
aorsf-fast	0.093 (0.003)	0.794 (0.004)	4.096	1.257
rsf-standard	0.091 (0.004)	0.785 (0.004)	7.684	0.943
aorsf-net	0.090 (0.003)	0.792 (0.004)	384.631	1.222
aorsf-cph	0.089 (0.002)	0.791 (0.005)	14.364	1.263
obliqueRSF-net	0.081 (0.002)	0.790 (0.005)	1858.149	373.179
glmnet-cox	0.077 (0.004)	0.785 (0.006)	1.828	0.017
nn-cox	0.073 (0.008)	0.779 (0.008)	31.694	72.618
cif-standard	0.073 (0.002)	0.787 (0.005)	70.222	351.070
ranger-extratrees	0.067 (0.003)	0.779 (0.008)	217.755	74.836
aorsf-random	0.059 (0.003)	0.751 (0.007)	9.041	1.119
cif-rotate	0.052 (0.002)	0.766 (0.008)	571.269	67.464
xgboost-cox	0.046 (0.010)	0.792 (0.003)	8.153	0.015

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
cif-extension	0.036 (0.001)	0.770 (0.007)	169.947	53.605
xgboost-aft	-6.55 (0.323)	0.791 (0.003)	31.120	0.013
Colon cancer; death, n = 929, p = 12				
aorsf-fast	0.100 (0.013)	0.719 (0.011)	0.236	0.051
cif-standard	0.100 (0.011)	0.713 (0.009)	0.663	3.547
aorsf-cph	0.098 (0.012)	0.717 (0.010)	0.640	0.056
aorsf-net	0.097 (0.010)	0.718 (0.009)	51.061	0.046
aorsf-random	0.095 (0.007)	0.717 (0.012)	1.038	0.046
cif-rotate	0.093 (0.015)	0.710 (0.012)	12.320	3.417
obliqueRSF-net	0.089 (0.005)	0.718 (0.010)	236.037	17.920
rsf-standard	0.088 (0.018)	0.703 (0.008)	1.281	0.150
cif-extension	0.082 (0.006)	0.711 (0.011)	7.857	3.918
ranger-extratrees	0.081 (0.008)	0.710 (0.012)	0.986	0.352
glmnet-cox	0.076 (0.013)	0.708 (0.018)	0.127	0.003
xgboost-cox	0.066 (0.008)	0.703 (0.011)	3.626	0.004
nn-cox	-0.004 (0.003)	0.481 (0.031)	15.080	1.153
xgboost-aft	-1.12 (0.267)	0.707 (0.013)	11.971	0.006
Colon cancer; recurrence, n = 929, p = 12				
aorsf-fast	0.097 (0.011)	0.714 (0.010)	0.232	0.051
aorsf-cph	0.097 (0.010)	0.713 (0.009)	0.651	0.051
cif-standard	0.094 (0.014)	0.704 (0.014)	0.662	3.381
aorsf-net	0.094 (0.015)	0.714 (0.012)	51.223	0.047
cif-rotate	0.092 (0.017)	0.700 (0.014)	11.852	3.577
obliqueRSF-net	0.087 (0.008)	0.711 (0.009)	247.097	15.943
rsf-standard	0.085 (0.016)	0.696 (0.011)	1.200	0.150
aorsf-random	0.083 (0.011)	0.702 (0.010)	0.958	0.046
cif-extension	0.081 (0.010)	0.708 (0.011)	8.680	4.310
ranger-extratrees	0.076 (0.014)	0.697 (0.012)	0.666	0.770
glmnet-cox	0.070 (0.013)	0.699 (0.016)	0.119	0.003
xgboost-cox	0.059 (0.012)	0.697 (0.013)	3.432	0.003
nn-cox	-0.008 (0.007)	0.524 (0.053)	13.774	0.938
xgboost-aft	-1.18 (0.255)	0.703 (0.014)	9.791	0.006
Early breast cancer; recurrence or death, n = 614, p = 1692				
obliqueRSF-net	0.064 (0.030)	0.739 (0.027)	1844.847	13.621

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
cif-rotate	0.061 (0.016)	0.735 (0.021)	7099.601	384.127
aorsf-cph	0.060 (0.043)	0.740 (0.020)	1.567	0.176
cif-standard	0.060 (0.020)	0.735 (0.024)	7.310	3.860
cif-extension	0.059 (0.016)	0.736 (0.025)	51.533	7.534
aorsf-fast	0.058 (0.041)	0.737 (0.023)	1.118	0.173
ranger-extratrees	0.053 (0.031)	0.730 (0.022)	0.387	1.014
xgboost-cox	0.034 (0.009)	0.736 (0.030)	2.480	0.007
glmnet-cox	0.026 (0.039)	0.711 (0.032)	5.909	0.006
aorsf-random	0.021 (0.016)	0.686 (0.029)	1.763	0.185
rsf-standard	0.021 (0.056)	0.695 (0.037)	0.356	0.625
aorsf-net	0.019 (0.095)	0.745 (0.022)	431.378	0.168
nn-cox	-0.024 (0.059)	0.656 (0.069)	17.924	1.897
xgboost-aft	-2.95 (0.868)	0.737 (0.022)	8.909	0.009
FCL; death, n = 541, p = 7				
glmnet-cox	0.122 (0.024)	0.801 (0.041)	0.259	0.002
aorsf-cph	0.104 (0.040)	0.775 (0.038)	0.165	0.021
aorsf-fast	0.101 (0.040)	0.774 (0.038)	0.083	0.021
aorsf-net	0.101 (0.040)	0.767 (0.039)	12.969	0.019
obliqueRSF-net	0.096 (0.028)	0.774 (0.038)	93.139	5.120
cif-standard	0.095 (0.039)	0.757 (0.039)	0.283	1.046
aorsf-random	0.093 (0.026)	0.766 (0.038)	0.293	0.020
cif-extension	0.090 (0.037)	0.737 (0.042)	4.384	2.202
cif-rotate	0.090 (0.050)	0.759 (0.031)	5.744	2.351
rsf-standard	0.081 (0.045)	0.737 (0.042)	0.109	0.042
ranger-extratrees	0.075 (0.012)	0.746 (0.038)	0.038	0.129
xgboost-cox	0.060 (0.060)	0.723 (0.118)	2.339	0.002
nn-cox	-0.005 (0.008)	0.530 (0.074)	12.863	0.443
xgboost-aft	-2.66 (0.438)	0.763 (0.047)	8.300	0.006
FCL; relapse, n = 541, p = 7				
glmnet-cox	0.031 (0.017)	0.625 (0.020)	0.270	0.003
xgboost-cox	0.014 (0.017)	0.610 (0.027)	1.867	0.003
ranger-extratrees	0.014 (0.019)	0.593 (0.031)	0.032	0.079
obliqueRSF-net	0.014 (0.015)	0.597 (0.027)	218.615	6.376
aorsf-random	0.011 (0.021)	0.601 (0.029)	0.428	0.021
aorsf-cph	0.009 (0.024)	0.600 (0.031)	0.276	0.026

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
aorsf-fast	0.007 (0.024)	0.599 (0.031)	0.116	0.024
aorsf-net	0.007 (0.026)	0.597 (0.033)	18.700	0.022
cif-standard	0.005 (0.020)	0.598 (0.026)	0.282	1.105
nn-cox	-0.002 (0.021)	0.556 (0.057)	11.608	0.458
cif-extension	-0.008 (0.026)	0.583 (0.034)	5.355	2.231
cif-rotate	-0.012 (0.025)	0.590 (0.033)	6.589	2.863
rsf-standard	-0.031 (0.035)	0.580 (0.032)	0.744	0.085
xgboost-aft	-0.915 (0.359)	0.584 (0.047)	6.551	0.006
GBSG II; recurrence or death, n = 686, p = 10				
obliqueRSF-net	0.123 (0.015)	0.744 (0.017)	284.837	7.018
cif-standard	0.123 (0.020)	0.743 (0.019)	0.425	1.696
rsf-standard	0.123 (0.022)	0.738 (0.018)	0.994	0.114
aorsf-net	0.121 (0.022)	0.738 (0.018)	38.168	0.038
aorsf-cph	0.118 (0.026)	0.733 (0.018)	0.405	0.041
aorsf-fast	0.115 (0.025)	0.731 (0.018)	0.165	0.041
cif-extension	0.115 (0.020)	0.744 (0.022)	8.087	3.238
cif-rotate	0.111 (0.022)	0.731 (0.015)	10.921	2.838
aorsf-random	0.103 (0.028)	0.721 (0.029)	0.755	0.037
ranger-extratrees	0.092 (0.024)	0.738 (0.032)	0.062	0.177
xgboost-cox	0.084 (0.014)	0.728 (0.018)	3.312	0.004
glmnet-cox	0.081 (0.019)	0.722 (0.022)	0.118	0.003
nn-cox	-0.004 (0.005)	0.509 (0.040)	14.255	1.400
xgboost-aft	-1.09 (0.122)	0.728 (0.019)	9.463	0.006
GUIDE-IT; CVD death, n = 894, p = 59				
aorsf-fast	0.077 (0.020)	0.746 (0.033)	0.176	0.039
aorsf-net	0.074 (0.019)	0.743 (0.033)	28.167	0.039
glmnet-cox	0.074 (0.041)	0.729 (0.085)	0.514	0.003
aorsf-cph	0.072 (0.021)	0.740 (0.038)	0.416	0.039
obliqueRSF-net	0.064 (0.013)	0.738 (0.028)	239.059	11.676
cif-rotate	0.062 (0.017)	0.715 (0.029)	34.699	4.860
cif-standard	0.059 (0.014)	0.732 (0.027)	0.790	3.278
cif-extension	0.053 (0.011)	0.724 (0.029)	14.497	6.192
ranger-extratrees	0.052 (0.014)	0.731 (0.031)	0.779	0.461
rsf-standard	0.047 (0.026)	0.709 (0.032)	0.177	0.059
aorsf-random	0.033 (0.010)	0.692 (0.035)	0.489	0.040

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
xgboost-cox	0.031 (0.052)	0.743 (0.024)	5.143	0.003
nn-cox	0.028 (0.038)	0.684 (0.085)	13.230	0.544
xgboost-aft	-5.44 (1.23)	0.729 (0.023)	13.003	0.006
GUIDE-IT; HF hospitalization, n = 894, p = 59				
aorsf-cph	0.083 (0.021)	0.724 (0.027)	0.707	0.057
aorsf-net	0.083 (0.020)	0.722 (0.024)	53.989	0.059
aorsf-fast	0.083 (0.022)	0.724 (0.026)	0.258	0.056
obliqueRSF-net	0.074 (0.013)	0.721 (0.024)	370.352	8.931
ranger-extratrees	0.073 (0.014)	0.720 (0.027)	0.501	0.191
cif-standard	0.069 (0.012)	0.714 (0.024)	0.795	3.266
cif-rotate	0.067 (0.023)	0.706 (0.031)	40.710	5.037
glmnet-cox	0.066 (0.022)	0.705 (0.026)	0.415	0.003
cif-extension	0.063 (0.010)	0.714 (0.019)	16.510	6.373
rsf-standard	0.055 (0.028)	0.693 (0.028)	1.136	0.117
aorsf-random	0.051 (0.013)	0.688 (0.028)	0.880	0.054
nn-cox	0.047 (0.032)	0.686 (0.061)	14.014	0.620
xgboost-cox	0.041 (0.013)	0.699 (0.028)	3.763	0.003
xgboost-aft	-2.23 (0.339)	0.695 (0.029)	16.064	0.007
Heart Transplant; graft-loss or death, n = 3787, p = 52				
cif-rotate	0.056 (0.012)	0.729 (0.022)	159.153	20.730
aorsf-fast	0.052 (0.008)	0.733 (0.020)	0.893	0.234
aorsf-cph	0.052 (0.008)	0.731 (0.018)	3.216	0.237
aorsf-net	0.052 (0.008)	0.730 (0.018)	123.347	0.229
rsf-standard	0.051 (0.009)	0.728 (0.014)	2.186	0.832
cif-standard	0.049 (0.008)	0.730 (0.018)	10.333	36.554
obliqueRSF-net	0.049 (0.007)	0.728 (0.020)	387.022	78.390
ranger-extratrees	0.045 (0.008)	0.727 (0.017)	5.104	6.168
glmnet-cox	0.035 (0.007)	0.707 (0.020)	1.413	0.006
cif-extension	0.035 (0.004)	0.725 (0.022)	50.118	18.785
xgboost-cox	0.029 (0.008)	0.715 (0.025)	4.244	0.007
nn-cox	0.029 (0.014)	0.693 (0.024)	14.198	8.076
aorsf-random	0.028 (0.003)	0.685 (0.013)	2.307	0.217
xgboost-aft	-4.31 (0.500)	0.720 (0.012)	13.609	0.007
MESA; coronary heart disease, n = 6785, p = 48				

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
aorsf-fast	0.064 (0.011)	0.807 (0.007)	1.306	0.374
aorsf-net	0.063 (0.012)	0.803 (0.008)	174.087	0.379
obliqueRSF-net	0.061 (0.008)	0.806 (0.007)	486.618	269.151
aorsf-cph	0.061 (0.011)	0.801 (0.010)	5.635	0.385
cif-rotate	0.058 (0.008)	0.801 (0.010)	311.704	46.721
cif-standard	0.058 (0.007)	0.801 (0.009)	24.315	98.498
rsf-standard	0.056 (0.014)	0.794 (0.011)	4.745	0.433
ranger-extratrees	0.047 (0.005)	0.795 (0.007)	6.566	11.394
cif-extension	0.047 (0.004)	0.803 (0.008)	103.940	34.786
glmnet-cox	0.040 (0.021)	0.774 (0.009)	5.519	0.006
nn-cox	0.032 (0.025)	0.760 (0.030)	22.091	19.964
aorsf-random	0.032 (0.006)	0.740 (0.015)	4.558	0.516
xgboost-cox	0.013 (0.019)	0.804 (0.007)	5.369	0.010
xgboost-aft	-9.73 (1.00)	0.802 (0.008)	26.660	0.011
MESA; death, n = 6793, p = 48				
aorsf-net	0.144 (0.009)	0.791 (0.010)	316.581	0.547
aorsf-fast	0.143 (0.009)	0.791 (0.011)	1.936	0.562
aorsf-cph	0.143 (0.008)	0.790 (0.010)	7.232	0.533
rsf-standard	0.140 (0.009)	0.783 (0.011)	3.983	0.470
obliqueRSF-net	0.138 (0.008)	0.789 (0.011)	1248.639	166.262
cif-standard	0.133 (0.007)	0.786 (0.011)	23.230	98.540
cif-rotate	0.124 (0.007)	0.779 (0.011)	322.280	37.520
glmnet-cox	0.114 (0.022)	0.782 (0.012)	1.774	0.007
ranger-extratrees	0.113 (0.005)	0.783 (0.010)	7.801	17.881
nn-cox	0.113 (0.041)	0.756 (0.091)	54.045	35.720
cif-extension	0.092 (0.003)	0.778 (0.010)	108.878	32.320
aorsf-random	0.070 (0.004)	0.727 (0.008)	7.162	0.550
xgboost-cox	0.057 (0.032)	0.791 (0.010)	10.747	0.010
xgboost-aft	-4.50 (0.243)	0.789 (0.009)	18.378	0.010
MESA; heart failure, n = 6785, p = 48				
aorsf-fast	0.114 (0.011)	0.866 (0.014)	1.172	0.324
aorsf-net	0.114 (0.011)	0.864 (0.016)	147.986	0.338
aorsf-cph	0.109 (0.010)	0.858 (0.017)	5.602	0.339
rsf-standard	0.108 (0.012)	0.857 (0.013)	2.784	0.895
obliqueRSF-net	0.107 (0.009)	0.868 (0.014)	391.939	332.468

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
cif-rotate	0.106 (0.011)	0.873 (0.014)	259.218	35.800
cif-standard	0.101 (0.010)	0.863 (0.015)	23.601	96.846
nn-cox	0.083 (0.015)	0.828 (0.024)	18.796	15.505
cif-extension	0.077 (0.005)	0.864 (0.012)	91.823	30.050
ranger-extratrees	0.075 (0.006)	0.847 (0.018)	7.714	11.655
aorsf-random	0.063 (0.006)	0.795 (0.019)	3.483	0.375
glmnet-cox	0.056 (0.047)	0.779 (0.148)	3.484	0.011
xgboost-cox	-0.008 (0.011)	0.873 (0.012)	8.137	0.010
xgboost-aft	-10.9 (1.08)	0.873 (0.015)	18.818	0.010
MESA; stroke, n = 6783, p = 48				
cif-rotate	0.025 (0.004)	0.766 (0.018)	269.642	35.908
aorsf-fast	0.025 (0.006)	0.766 (0.015)	1.144	0.325
glmnet-cox	0.025 (0.006)	0.769 (0.017)	3.082	0.006
obliqueRSF-net	0.024 (0.004)	0.767 (0.016)	357.106	281.765
cif-standard	0.024 (0.004)	0.763 (0.018)	24.960	98.465
aorsf-net	0.024 (0.006)	0.759 (0.016)	138.120	0.340
aorsf-cph	0.024 (0.005)	0.759 (0.016)	4.983	0.325
ranger-extratrees	0.021 (0.003)	0.758 (0.018)	7.114	11.973
cif-extension	0.021 (0.002)	0.772 (0.019)	95.005	32.157
rsf-standard	0.019 (0.010)	0.745 (0.020)	2.679	0.865
nn-cox	0.017 (0.005)	0.732 (0.037)	18.484	20.376
aorsf-random	0.013 (0.003)	0.713 (0.026)	3.517	0.343
xgboost-cox	0.003 (0.021)	0.765 (0.023)	5.210	0.009
xgboost-aft	-13.9 (1.53)	0.769 (0.018)	17.608	0.009
Monoclonal gammopathy; death, n = 1384, p = 8				
aorsf-cph	0.157 (0.017)	0.741 (0.012)	1.325	0.103
cif-rotate	0.156 (0.023)	0.742 (0.018)	15.136	4.979
aorsf-fast	0.155 (0.017)	0.741 (0.012)	0.434	0.097
obliqueRSF-net	0.153 (0.015)	0.742 (0.014)	231.507	12.455
aorsf-net	0.152 (0.017)	0.739 (0.012)	90.204	0.085
cif-standard	0.148 (0.017)	0.736 (0.013)	0.928	5.907
rsf-standard	0.148 (0.018)	0.735 (0.012)	1.988	0.193
aorsf-random	0.145 (0.014)	0.734 (0.013)	1.821	0.085
cif-extension	0.141 (0.010)	0.744 (0.015)	9.939	4.749
glmnet-cox	0.129 (0.019)	0.722 (0.014)	0.308	0.003

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
xgboost-cox	0.123 (0.012)	0.731 (0.014)	5.767	0.005
ranger-extratrees	0.115 (0.005)	0.742 (0.014)	0.062	0.364
nn-cox	0.062 (0.071)	0.632 (0.118)	17.238	0.720
xgboost-aft	-0.870 (0.112)	0.732 (0.016)	15.477	0.007
Monoclonal gammopathy; malignancy, n = 1384, p = 8				
glmnet-cox	0.018 (0.010)	0.674 (0.032)	0.324	0.002
aorsf-cph	0.011 (0.009)	0.646 (0.037)	0.650	0.044
xgboost-cox	0.010 (0.007)	0.647 (0.042)	1.962	0.003
aorsf-fast	0.010 (0.009)	0.642 (0.038)	0.194	0.044
obliqueRSF-net	0.008 (0.006)	0.635 (0.032)	39.065	15.966
ranger-extratrees	0.007 (0.007)	0.644 (0.035)	0.061	0.197
cif-extension	0.007 (0.008)	0.629 (0.028)	8.637	4.451
cif-standard	0.007 (0.007)	0.630 (0.032)	0.956	6.010
aorsf-net	0.006 (0.010)	0.645 (0.036)	23.264	0.042
aorsf-random	0.006 (0.014)	0.639 (0.037)	0.519	0.042
nn-cox	-0.005 (0.006)	0.528 (0.083)	10.752	0.593
rsf-standard	-0.009 (0.015)	0.626 (0.032)	0.634	0.069
cif-rotate	-0.034 (0.022)	0.545 (0.032)	12.464	4.154
xgboost-aft	-5.78 (1.22)	0.627 (0.045)	9.325	0.006
Movies released in 2015-2018; gross 1M USD, n = 551, p = 46				
cif-rotate	0.645 (0.026)	0.945 (0.008)	20.183	3.526
glmnet-cox	0.629 (0.037)	0.940 (0.011)	0.198	0.003
aorsf-net	0.535 (0.033)	0.929 (0.012)	50.651	0.044
aorsf-cph	0.527 (0.024)	0.926 (0.013)	0.792	0.045
rsf-standard	0.525 (0.023)	0.923 (0.013)	0.746	0.105
aorsf-fast	0.521 (0.029)	0.923 (0.015)	0.218	0.044
xgboost-cox	0.515 (0.029)	0.933 (0.011)	14.297	0.005
nn-cox	0.497 (0.143)	0.889 (0.056)	16.418	0.705
cif-standard	0.472 (0.030)	0.900 (0.021)	0.379	1.292
cif-extension	0.449 (0.026)	0.919 (0.015)	8.850	3.931
ranger-extratrees	0.432 (0.019)	0.896 (0.020)	0.070	0.127
obliqueRSF-net	0.318 (0.025)	0.908 (0.021)	152.602	9.347
aorsf-random	0.293 (0.041)	0.841 (0.034)	0.999	0.041
xgboost-aft	-0.435 (0.057)	0.930 (0.011)	28.847	0.007

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
Non-alcohol fatty liver disease; death, n = 17549, p = 24				
aorsf-cph	0.214 (0.009)	0.871 (0.005)	20.799	1.406
aorsf-fast	0.214 (0.010)	0.871 (0.006)	4.828	1.422
aorsf-net	0.212 (0.009)	0.867 (0.006)	446.975	1.285
obliqueRSF-net	0.210 (0.008)	0.870 (0.006)	1520.102	1097.379
glmnet-cox	0.209 (0.010)	0.861 (0.005)	2.153	0.014
rsf-standard	0.207 (0.009)	0.860 (0.006)	11.043	1.209
cif-standard	0.206 (0.007)	0.864 (0.007)	65.607	616.634
cif-rotate	0.191 (0.006)	0.866 (0.005)	256.009	62.262
ranger-extratrees	0.182 (0.007)	0.862 (0.005)	40.654	303.362
cif-extension	0.167 (0.003)	0.868 (0.006)	122.213	55.889
aorsf-random	0.142 (0.007)	0.843 (0.006)	11.433	1.500
xgboost-cox	0.020 (0.015)	0.876 (0.005)	11.932	0.021
nn-cox	-0.001 (0.003)	0.580 (0.090)	21.950	95.510
xgboost-aft	-7.74 (0.684)	0.875 (0.006)	33.565	0.015
Primary biliary cholangitis; death, n = 276, p = 19				
aorsf-fast	0.436 (0.026)	0.909 (0.020)	0.071	0.021
aorsf-cph	0.423 (0.027)	0.908 (0.019)	0.152	0.020
aorsf-net	0.415 (0.031)	0.907 (0.020)	13.638	0.018
cif-rotate	0.404 (0.043)	0.901 (0.022)	9.187	2.214
rsf-standard	0.397 (0.023)	0.895 (0.020)	0.087	0.036
obliqueRSF-net	0.366 (0.034)	0.908 (0.020)	106.606	1.721
aorsf-random	0.353 (0.028)	0.895 (0.020)	0.288	0.019
cif-standard	0.344 (0.034)	0.904 (0.023)	0.185	0.362
glmnet-cox	0.343 (0.059)	0.894 (0.024)	0.120	0.003
cif-extension	0.341 (0.032)	0.904 (0.019)	5.195	2.236
ranger-extratrees	0.267 (0.029)	0.897 (0.023)	0.027	0.037
xgboost-cox	0.250 (0.108)	0.881 (0.027)	6.107	0.003
nn-cox	-0.028 (0.019)	0.652 (0.123)	9.552	0.240
xgboost-aft	-0.905 (0.366)	0.885 (0.021)	8.233	0.005
Rotterdam tumor bank; death, n = 2982, p = 11				
aorsf-net	0.165 (0.010)	0.760 (0.008)	148.475	0.186
obliqueRSF-net	0.162 (0.009)	0.760 (0.009)	439.574	36.953
aorsf-cph	0.161 (0.010)	0.758 (0.009)	2.601	0.205
aorsf-fast	0.159 (0.011)	0.756 (0.009)	0.809	0.206

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
cif-standard	0.159 (0.008)	0.759 (0.008)	4.308	23.147
rsf-standard	0.157 (0.011)	0.755 (0.008)	2.754	0.522
aorsf-random	0.153 (0.007)	0.752 (0.009)	2.869	0.184
cif-rotate	0.151 (0.012)	0.754 (0.013)	33.457	8.469
ranger-extratrees	0.139 (0.005)	0.749 (0.009)	2.721	3.105
xgboost-cox	0.134 (0.011)	0.754 (0.010)	4.785	0.005
cif-extension	0.130 (0.005)	0.751 (0.009)	24.345	9.889
glmnet-cox	0.120 (0.007)	0.733 (0.010)	0.499	0.004
nn-cox	-0.028 (0.087)	0.535 (0.057)	16.443	7.747
xgboost-aft	-1.37 (0.126)	0.759 (0.008)	14.195	0.007
Rotterdam tumor bank; recurrence, n = 2982, p = 11				
obliqueRSF-net	0.148 (0.011)	0.738 (0.011)	513.045	37.957
aorsf-net	0.147 (0.012)	0.736 (0.011)	169.714	0.197
aorsf-cph	0.145 (0.012)	0.735 (0.010)	2.867	0.216
cif-standard	0.144 (0.012)	0.735 (0.011)	4.369	23.006
aorsf-fast	0.143 (0.011)	0.734 (0.010)	0.896	0.218
aorsf-random	0.140 (0.010)	0.731 (0.010)	3.326	0.191
rsf-standard	0.138 (0.011)	0.732 (0.009)	2.809	0.281
ranger-extratrees	0.135 (0.008)	0.735 (0.011)	2.807	2.778
cif-rotate	0.133 (0.009)	0.728 (0.010)	35.059	8.648
cif-extension	0.119 (0.007)	0.733 (0.010)	22.197	8.887
glmnet-cox	0.118 (0.007)	0.728 (0.011)	0.498	0.004
xgboost-cox	0.114 (0.007)	0.730 (0.012)	4.233	0.005
nn-cox	-0.002 (0.002)	0.491 (0.028)	24.631	11.769
xgboost-aft	-1.06 (0.147)	0.735 (0.011)	13.403	0.007
Serum free light chain; death, n = 7874, p = 10				
aorsf-fast	0.256 (0.016)	0.826 (0.010)	2.468	0.600
aorsf-cph	0.255 (0.016)	0.826 (0.011)	7.124	0.601
aorsf-net	0.254 (0.014)	0.824 (0.011)	274.793	0.556
obliqueRSF-net	0.251 (0.014)	0.822 (0.011)	1071.189	145.107
glmnet-cox	0.251 (0.015)	0.821 (0.010)	0.546	0.011
rsf-standard	0.247 (0.016)	0.817 (0.011)	5.052	0.549
cif-standard	0.247 (0.014)	0.819 (0.011)	18.346	116.811
ranger-extratrees	0.247 (0.012)	0.822 (0.010)	7.234	16.737
aorsf-random	0.236 (0.014)	0.819 (0.011)	6.056	0.566

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
cif-rotate	0.228 (0.010)	0.819 (0.009)	62.380	20.093
cif-extension	0.202 (0.005)	0.821 (0.010)	43.862	21.183
xgboost-cox	0.104 (0.034)	0.824 (0.011)	6.743	0.009
nn-cox	0.001 (0.003)	0.610 (0.135)	23.112	21.723
xgboost-aft	-2.75 (0.305)	0.824 (0.011)	30.052	0.009
SPRINT; CVD death, n = 9361, p = 174				
glmnet-cox	0.070 (0.007)	0.792 (0.008)	12.779	0.016
aorsf-cph	0.067 (0.005)	0.791 (0.010)	10.753	0.649
aorsf-net	0.067 (0.005)	0.788 (0.010)	339.857	0.652
aorsf-fast	0.067 (0.004)	0.791 (0.010)	2.950	0.653
obliqueRSF-net	0.066 (0.004)	0.790 (0.011)	1071.167	425.299
rsf-standard	0.062 (0.006)	0.780 (0.015)	4.271	0.599
cif-standard	0.060 (0.003)	0.791 (0.010)	47.426	180.930
cif-rotate	0.059 (0.005)	0.783 (0.008)	941.817	112.906
ranger-extratrees	0.053 (0.003)	0.786 (0.013)	5.733	9.972
nn-cox	0.042 (0.013)	0.765 (0.016)	22.162	29.154
cif-extension	0.033 (0.002)	0.782 (0.008)	121.313	32.071
aorsf-random	0.025 (0.002)	0.736 (0.013)	5.621	0.772
xgboost-cox	0.011 (0.017)	0.793 (0.009)	8.908	0.014
xgboost-aft	-10.7 (0.681)	0.789 (0.011)	24.248	0.015
SPRINT; death, n = 9361, p = 174				
glmnet-cox	0.123 (0.008)	0.770 (0.010)	4.957	0.011
aorsf-cph	0.116 (0.006)	0.769 (0.008)	14.405	1.433
aorsf-fast	0.115 (0.007)	0.768 (0.007)	4.124	1.494
aorsf-net	0.112 (0.007)	0.766 (0.008)	591.726	0.942
obliqueRSF-net	0.111 (0.006)	0.766 (0.008)	2920.730	233.477
rsf-standard	0.109 (0.006)	0.760 (0.007)	6.036	0.711
cif-standard	0.105 (0.006)	0.763 (0.009)	48.060	191.307
nn-cox	0.098 (0.010)	0.752 (0.013)	29.086	28.125
ranger-extratrees	0.096 (0.005)	0.755 (0.008)	10.665	22.541
cif-rotate	0.091 (0.005)	0.744 (0.008)	1083.768	118.583
cif-extension	0.055 (0.002)	0.745 (0.008)	135.963	32.414
aorsf-random	0.051 (0.004)	0.717 (0.010)	9.053	1.029
xgboost-cox	0.034 (0.025)	0.771 (0.008)	12.900	0.015
xgboost-aft	-4.49 (0.313)	0.771 (0.007)	23.081	0.013

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
Systolic Heart Failure; death, n = 2231, p = 41				
glmnet-cox	0.117 (0.011)	0.747 (0.013)	0.304	0.004
obliqueRSF-net	0.115 (0.010)	0.747 (0.010)	376.210	26.092
cif-rotate	0.114 (0.011)	0.741 (0.011)	70.159	10.535
aorsf-net	0.113 (0.012)	0.744 (0.012)	121.294	0.159
aorsf-cph	0.113 (0.012)	0.745 (0.011)	1.987	0.150
aorsf-fast	0.111 (0.015)	0.744 (0.011)	0.616	0.151
cif-standard	0.110 (0.009)	0.744 (0.010)	3.623	15.504
rsf-standard	0.107 (0.011)	0.737 (0.010)	2.132	0.244
cif-extension	0.095 (0.004)	0.744 (0.011)	28.859	9.866
xgboost-cox	0.092 (0.007)	0.746 (0.010)	5.858	0.006
ranger-extratrees	0.090 (0.008)	0.734 (0.012)	2.841	1.923
aorsf-random	0.080 (0.004)	0.727 (0.009)	2.143	0.147
nn-cox	0.072 (0.029)	0.705 (0.034)	33.965	7.798
xgboost-aft	-2.02 (0.169)	0.741 (0.010)	15.753	0.007
VA lung cancer trial; death, n = 137, p = 8				
cif-rotate	0.197 (0.066)	0.791 (0.036)	5.405	1.442
aorsf-fast	0.191 (0.049)	0.792 (0.039)	0.052	0.013
aorsf-cph	0.190 (0.055)	0.793 (0.043)	0.100	0.014
aorsf-net	0.189 (0.050)	0.793 (0.041)	9.623	0.012
rsf-standard	0.162 (0.048)	0.777 (0.046)	0.069	0.026
cif-extension	0.160 (0.052)	0.787 (0.039)	3.746	1.229
glmnet-cox	0.145 (0.031)	0.790 (0.033)	0.286	0.002
aorsf-random	0.143 (0.052)	0.774 (0.046)	0.244	0.012
cif-standard	0.121 (0.041)	0.769 (0.039)	0.096	0.122
obliqueRSF-net	0.117 (0.038)	0.796 (0.028)	60.816	0.674
ranger-extratrees	0.079 (0.039)	0.775 (0.044)	0.022	0.028
xgboost-cox	0.027 (0.068)	0.725 (0.046)	1.223	0.003
xgboost-aft	-0.011 (0.112)	0.757 (0.053)	5.323	0.006
nn-cox	-0.037 (0.043)	0.555 (0.067)	10.089	0.118

References

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*, 32(30):5381–5397, 2013.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2022. URL `https://mc-stan.org/rstanarm/`. R package version 2.21.3.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999. URL `https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5`.

Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030. URL `https://doi.org/10.1001/jama.1982.03320430047030`.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23(1):77–91, 2004.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15 (3):651–674, 2006.

Torsten Hothorn, Kurt Hornik, Carolin Strobl, and Achim Zeileis. Party: a laboratory for recursive partytioning, 2010.

H. Ishwaran and U.B. Kogalur. *Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2019. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 2.8.0, available at `https://cran.r-project.org/package=randomForestSRC`.

Hemant Ishwaran and Udaya B Kogalur. Consistency of random survival forests. *Statistics & probability letters*, 80(13-14):1056–1064, 2010.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

Byron C Jaeger, D Leann Long, Dustin M Long, Mario Sims, Jeff M Szychowski, Yuan-I Min, Leslie A Mcclure, George Howard, and Noah Simon. Oblique random survival forests. *The Annals of Applied Statistics*, 13(3):1847–1883, 2019.

Michael W Kattan and Thomas A Gerds. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and prognostic research*, 2(1):1–7, 2018.

Rakesh Katuwal, Ponnuthurai Nagaratnam Suganthan, and Le Zhang. Heterogeneous oblique random forest. *Pattern Recognition*, 99:107078, 2020.

Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL `https://www.tidymodels.org`.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Bjoern H Menze, B Michael Kelm, Daniel N Splitthoff, Ullrich Koethe, and Fred A Hamprecht. On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 453–469. Springer, 2011.

Karel GM Moons, Andre Pascal Kengne, Diederick E Grobbee, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Mark Woodward. Risk prediction models: II. external validation, model updating, and impact assessment. *Heart*, 98(9):691–698, 2012a.

Karel GM Moons, Andre Pascal Kengne, Mark Woodward, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Diederick E Grobbee. Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, 98(9):683–690, 2012b.

Nitesh Poona, Adriaan Van Niekerk, and Riyad Ismail. Investigating the utility of oblique tree-based ensembles for the classification of hyperspectral data. *Sensors*, 16(11):1918, 2016.

Xueheng Qiu, Le Zhang, Ponnuthurai Nagaratnam Suganthan, and Gehan AJ Amaratunga. Oblique random forest ensemble via least square estimation for time series forecasting. *Information Sciences*, 420:249–262, 2017.

Tom Rainforth and Frank Wood. Canonical correlation forests. *arXiv preprint arXiv:1507.05444*, 2015.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

Terry Therneau. Survival package source code documentation, April 2022. URL https://github.com/therneau/survival/blob/5440691d44abea537b08aeb60153a31654d66a9b/noweb. original-date: 2016-04-28.

Tyler M Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse L Patsolic, Benjamin Falk, Carey E Priebe, Jason Yim, Randal Burns, Mauro Maggioni, et al. Sparse projection oblique randomer forests. *Journal of machine learning research*, 21(104), 2020.

Hong Wang and Gang Li. A selective review on random survival forests for high dimensional data. *Quantitative bio-science*, 36(2):85, 2017.

Hong Wang and Lifeng Zhou. Random survival forest with space extensions for censored data. *Artificial intelligence in medicine*, 79:52–61, 2017.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01. URL https://www.jstatsoft.org/index.php/jss/article/view/v077i01.

Le Zhang and Ponnuthurai N Suganthan. Oblique decision tree ensemble via multisurface proximal support vector machine. *IEEE transactions on cybernetics*, 45(10):2165–2176, 2014.

Lifeng Zhou, Hong Wang, and Qingsong Xu. Random rotation survival forest for high dimensional censored data. *SpringerPlus*, 5(1):1–10, 2016.

Ruoqing Zhu. *Tree-based Methods for Survival Analysis and High-dimensional Data*. PhD thesis, The University of North Carolina at Chapel Hill, 2013.

Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.